Estimation of the Prevalence of Very Rare Diseases Based on Data From Specialized Treatment Centers: Approaches for the Identification of the Reference Population

Dana Y. Teltsch, PhD, Epidemiologist, Evidera
Richard Scott Swain, MPH, Research Associate, Evidera
Matthew W. Reynolds, PhD, Vice President, Epidemiology, Evidera

Studies of very rare diseases (less than 1 in 100,000 of the general population) often use cases seen at specialized centers. While the estimation of the disease prevalence based on data from such studies is typically complicated by multiple potential sources of both systematic error and random error, establishing the reference population is a key challenge. We discuss several approaches for the estimation of the reference population and give examples based on a study of Multicentric Castleman Disease (MCD).

FACTORS THAT COMPLICATE PREVALENCE ESTIMATION OF RARE DISEASES

- In rare diseases with non-simple diagnoses, a large number of cases are undiagnosed, or diagnosed with great delay. The true number of cases in the population is likely higher.
- Often diseases were not well studied; diseases' natural history and duration are not well-known.
- Only a small number of centers and patients are available for study.
- Centers that treat a relatively large number of patients are specialized centers or known centers of excellence and serve as referral rather than regional centers. There is not, therefore, a well-defined geographical area where patients are coming from.
- Patients will travel long distances, or even relocate to seek treatment.
- Diseases may be related with certain ethnic or racial backgrounds, environmental, occupational or behavioral factors. These may be associated with geographical areas and vary by the location of centers and complicate generalization of prevalence estimates.
- It might be impossible to distinguish true incidence and prevalence from referral patterns or access to care.

STUDY EXAMPLE — MCD

Multicentric Castleman Disease (MCD) is a rare lymphoproliferative disease with no established therapy and of unknown origin that involves the overproduction of the cytokine interleukin-6 as one of the key pathogenic processes.[1] MCD patients are often heterogeneous in signs and symptoms, some of the more frequent being fatigue, night sweats, fevers and anemia. Chronic therapy and optimal disease control are the present clinical practice and goal, respectively.[2,3]

Adult patients with a confirmed MCD diagnosis between Jan. 1, 2000, and Dec. 31, 2009, from two major referral centers that specialize in treating MCD — Mayo Clinic (Mayo Clinic; Rochester, MN) and the Fred Hutchinson Cancer Research Center (FHCRC; Seattle, WA) — were included, and their electronic medical records were abstracted. One of the study objectives was the estimation of the disease prevalence.

Assessment of the Reference Population through Catchment Area

The catchment area defines the area from which patients will most likely be referred to the specific center and, therefore, included in the data. The reference population for each center can therefore be assumed to compose the residents of the catchment area. The reference population can be estimated using U.S. Census data.

In our study, analyses were performed using ArcGIS and Census 2007 data. Stratification by age, sex, race, ethnicity and educational attainment was based on the Census 2000 data.

The maps in the figures display the location of MCD cases identified by the two centers and catchment

areas assessed through different approaches. Cases for each center are represented by a dot. The location for the patient was available only as the 3-digit ZIP code area they resided in at the time of diagnosis. Therefore, the locations of the dots displayed on the maps were randomly placed within each representative 3-digit ZIP code area by ArcMap.

We assumed that the changes in the states' population over the six years prior to 2007 and the two years post 2007 were not significant for the estimate. Generalization of the results to estimate the national prevalence of the disease will have to either assume that changes over the study time to local populations were similar to the national ones, or take them into account in the calculations.

In addition, the two centers are well-known centers of excellence, and patients might not represent the general patient population. Differences in distributions of risk factors such as gender, age, HIV status and other disease risk factors might vary between the centers' population and the general population, complicating the generalization of the estimates.

Assessment of the Catchment Area

Spatial Distribution of Cases in the MCD Study

The Mayo Clinic seems to serve as a referral center, with cases originating from a vast geographic area *(Figure 2)*, including two cases from Washington state. The Mayo Clinic cases were based in ZIP codes from 16 states, with no state represented by more than five patients.

Most patients from the FHCRC center were located in ZIP code areas in Washington state and Oregon *(Figure 2)*. One patient from the FHCRC center with a Washington state area code did not have a ZIP code available and he was assigned to the most common ZIP code.

Regional-Based Catchment Definition

Regional-based catchment definitions could be based on observed spatial patterns in the data or on information about referral patterns from the institute or other sources.

The catchment area for the FHCRC was defined based on clustered MCD cases in the states of Washington and Oregon. These states also had by far the highest prevalence proportion at one to two MCD cases per million population *(Figure 1).* The FHCRC did not catch all MCD cases from within Washington and Oregon (in fact two cases in this area were identified by the Mayo Clinic); however, the spatial clustering of cases in these two states is reasonable justification for the definition of the catchment area. A decreasing gradient with distance along the West Coast was apparent. Washington state cases and population can therefore be used, by this approach, as the basis for the prevalence estimate. Many cases are likely not represented in the data, even within the catchment area (by any definition), and due to the difficulty of diagnosis, and possibly lack of access to care, many MCD patients are likely never diagnosed. The estimates are therefore best used as a lower limit to the likely true number of MCD cases, and the estimates based on areas with higher prevalence are likely closer to the true prevalence.

Driving-Distance-Based Catchment Areas Definition

Catchment areas based on driving distances by categories are presented in *Figure 2*. Thresholds could be chosen by assumptions regarding the time period most patients would be willing to travel.

Cases-Clustering-Based Catchment Areas

We used the "Hotspot Analysis" tool in ArcMap to define catchment areas for each center based off of the 10-year period prevalence proportion for each 3-digit ZIP code area. Spatial relationship was based on inverse distance squared (strong punishment for increasing distance,

Figure 1

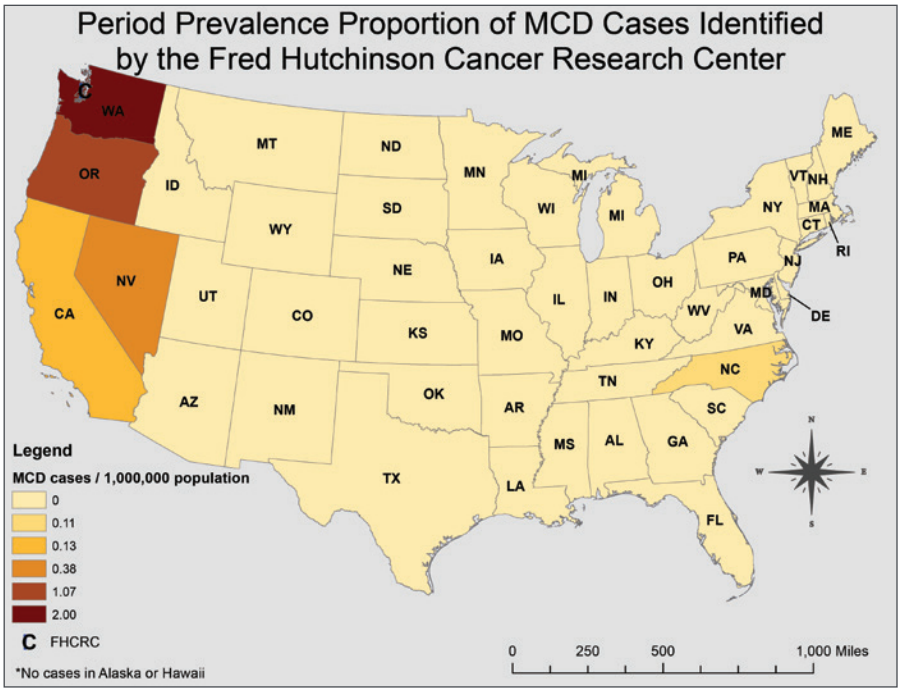
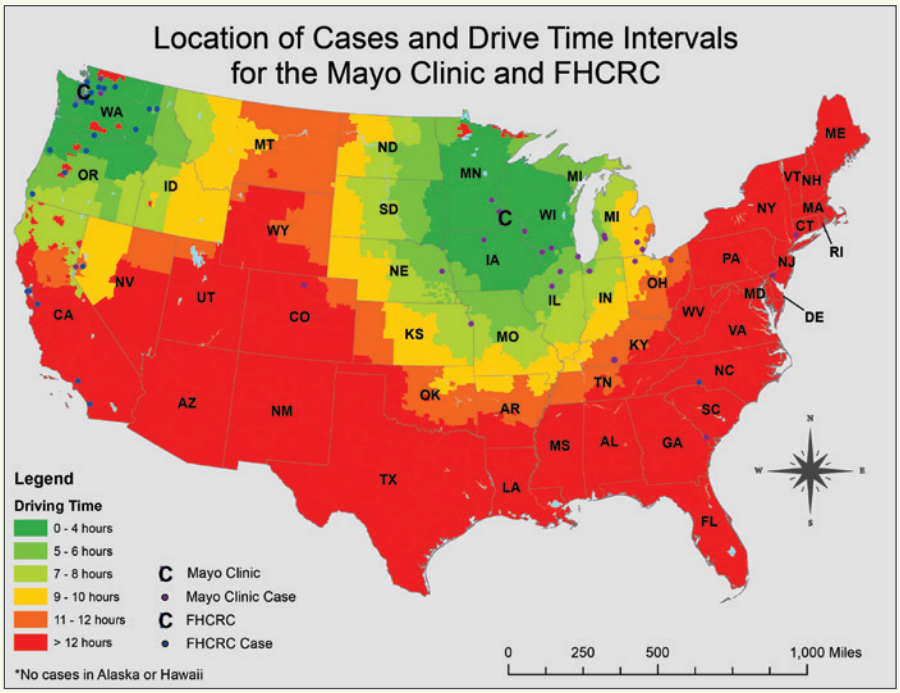


Figure 2



as we believed increasing driving distance to the center would be a barrier to treatment). The distance method used was the Manhattan distance that accounts for people traveling by roadways.

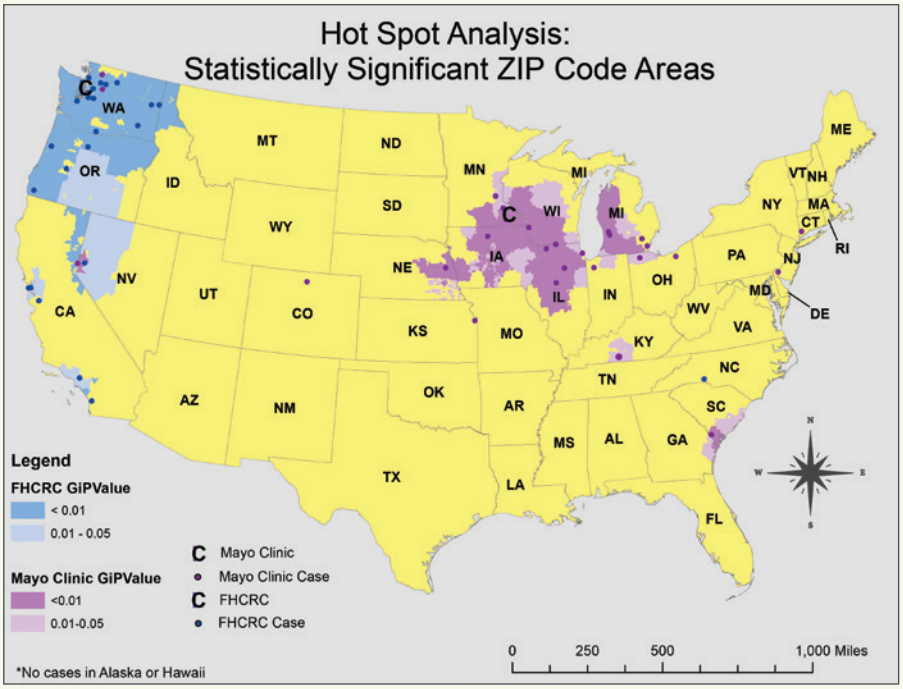
ArcMap uses this information to generate a Z-score. The significance level of the Z-score (areas where Z>1.96; p<0.05 indicate a cluster) is displayed in *Figure 3*. We considered the contiguous cluster around each center to be the catchment area.

The 10-year period prevalence proportion was calculated for each 3-digit ZIP code area by dividing the number of cases by the total population estimated by Census 2007 data (period prevalence per million population = MCD cases / 2007 population × 1,000,000).

Results from the broader study of patients' education level and their location indirectly supported this definition of catchment area. We compared patients' education grouped into two levels for Mayo Clinic patients (for which education-level information was available for all patients). Education of the adult population was compared by location within and outside of the catchment area. A significantly higher percentage of patients with higher levels of education (graduate/professional degree or higher) compared to the general adult population in the area traveled from outside of the catchment area to receive care at the center. This was not the case for patients with a lower level of education, and was not the case for patients with higher education within the catchment area. These results could suggest that broad socio-economic strata were using the Mayo Clinic for their care, whereas those from more distal locations tended to be from higher education (and likely income) strata.

A SIGNIFICANTLY HIGHER PERCENTAGE OF PATIENTS WITH HIGHER LEVELS OF EDUCATION (GRADUATE/ PROFESSIONAL DEGREE OR HIGHER) COMPARED TO THE GENERAL ADULT POPULATION IN THE AREA TRAVELED FROM OUTSIDE OF THE CATCHMENT AREA TO RECEIVE CARE AT THE CENTER.

Figure 3

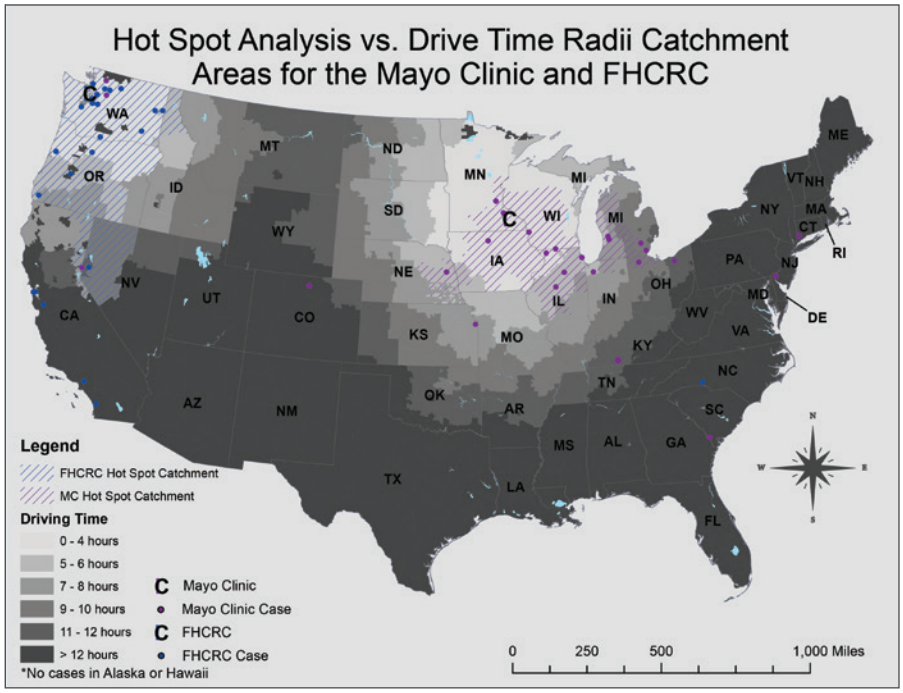


SUMMARY

There is significant overlap among the catchment areas defined by the different methods (*Figure 4*).

The "hot spots" based catchment areas at a 0.05 significance level are influenced by the population density in an area, and therefore areas that are sparsely populated but with close proximity to a center, or in certain geographical areas, might not be included in catchment areas by this approach.

Figure 4



EVEN WITHIN THE SAME CENTER, CATCHMENT AREAS MAY DIFFER FOR DIFFERENT DISEASES ACCORDING TO DISEASE RARITY, IMPACT ON PATIENTS' LIVES, REPUTATION OF THE CENTER AND OTHER FACTORS.

The most appropriate choice would depend on the study design and objective and on the data. The regional-based approach is the easiest to implement and could offer a simple solution for a rough estimate. The choice of approach for the estimation of a catchment area should also be determined by the characteristics of the disease in question and of the participating centers. Even within the same center, catchment areas may differ for different diseases according to disease rarity, impact on patients' lives, reputation of the center and other factors.

For more information, please contact Dana.Teltsch@evidera.com, Richard.Swain@evidera.com or Matthew.Reynolds@evidera.com.

References

[1] Brandt SJ, Bodine DM, Dunbar CE, Nienhuis AW. Dysregulated Interleukin 6 Expression Produces a Syndrome Resembling Castleman's Disease in Mice. *J Clin Invest.* 1990 Aug; 86(2):592-599.

[2] Casper C. The Aetiology and Management of Castleman Disease at 50 years: Translating Pathophysiology to Patient Care. *Br J Haematol*.2005 Apr; 129(1):3-17.

[3] van Rhee F, Stone K, Szmania S, Barlogie B, Singh Z. Castleman Disease in the 21st Century: An Update on Diagnosis, Assessment, and Therapy. *Clin Adv Hematol Oncol.* 2010 July; 8(7):486-498.

[4] Robinson D Jr, Reynolds M, Casper C, Dispenzieri A, Vermeulen J, Payne K, Schramm J, Ristow K, Desrosiers MP, Yeomans K, Teltsch D, Swain R, Habermann TM, Rotella P, Van de Velde H.Clinical Epidemiology and Treatment Patterns of Patients with Multicentric Castleman Disease: Results from Two US Treatment Centres. *Br J Haematol.* 2014 Apr; 165(1):39-48. Epub 2014 Jan 6.